Interspeech 2022
18-22 September 2022, Incheon, Korea

An Empirical Analysis on the Vulnerabilities of End-to-End Speech Segregation Models

Rahil Parikh[1], *Gaspar Rochette*[2], *Carol Espy-Wilson*[1], *Shihab Shamma*[1]

[1]Institute for Systems Research
University of Maryland College Park, USA
[2]ENS Paris, PSL Université, France

rahil@umd.edu

Abstract

End-to-end learning models have demonstrated a remarkable capability in performing speech segregation. Despite their wide-scope of real-world applications, little is known about the mechanisms they employ to group and consequently segregate individual speakers. Knowing that harmonicity is a critical cue for these networks to group sources [1], in this work, we perform a thorough investigation on ConvTasnet [2] and DPT-Net [3] to analyze how they perform a harmonic analysis of the input mixture. We perform ablation studies where we apply low-pass, high-pass, and band-stop filters of varying pass-bands to empirically analyze the harmonics most critical for segregation. We also investigate how these networks decide which output channel to assign to an estimated source by introducing discontinuities in synthetic mixtures. We find that end-to-end networks are highly unstable, and perform poorly when confronted with deformations which are imperceptible to humans. Replacing the encoder in these networks with a spectrogram leads to lower overall performance, but much higher stability. This work helps us to understand what information these network rely on for speech segregation, and exposes two sources of generalization-errors. It also pinpoints the encoder as the part of the network responsible for these errors, allowing for a redesign with expert knowledge or transfer learning.

Index Terms: end-to-end speech segregation, Conv-Tasnet, harmonic continuity, temporal consistency, generalization

1. Introduction

Speech communication often occurs in complex acoustic environments with concurrent sounds. To realize speech processing applications such as Automatic Speech Recognition, Speaker Verification and Identification, etc. in such environments, the system must be capable of segregating speech of individual speakers from a mixture of speakers. Advances in deep learning have facilitated drastic improvements in single-channel speech segregation [2–13]. Some of these models [2, 6] outperform the Ideal Ratio Mask (IRM) [14] when trained and evaluated on the WSJ-2-Mix dataset [4], allowing them to be used in a wide variety of downstream applications in the *wild*. These advances have been adopted to other source segregation tasks such as Universal Source Segregation [15] and Music Segregation [16], directly affecting an even wider array of downstream audio tasks, such as Sound Event Detection [17, 18].

End-to-end (E2E) speech-segregation systems are believed to out-perform spectrogram based models [6]. A typical design is shown in Fig 1, where the model is trained to segregate a mixture of C speakers into C individual streams. The separation sub-net generates C masks on the encoded mixture waveform. The element-wise product of the encoded mixture and masks

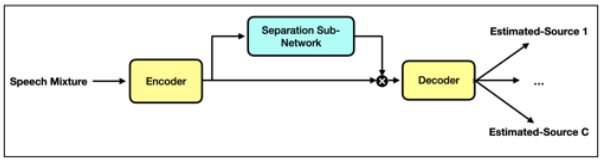

Figure 1: *A typical E2E speech segregation network design*

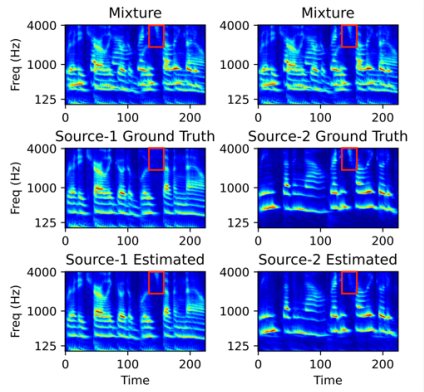

Figure 2: *ConvTasnet fails to segregate fricatives in natural speech. The red boxes highlight where '/s/' is grouped with Source-1 instead of Source-2.*

are decoded to produce each estimated speaker. The encoder, separation sub-network and decoder are trained end-to-end to optimize the Scale Invariant Source-to-Noise Ratio (SI-SNR) or Signal-to-Distortion Ratio (SDR) [19–21] using Permutation Invariant Training (PIT) [8].

Due to the black-box nature of deep-learning models, little is known about the principles employed by these models to segregate sources. The sensitivity of E2E speech-segregation models to mixtures of inharmonic speech is demonstrated in [1, 22], indicating that these networks are heavily reliant on harmonicity to group and segregate sources. This is corroborated by our observation that despite the abundance of fricatives in the training data, ConvTasnet often fails to segregate fricatives in natural speech as demonstrated in Fig 2, although, this error is usually not heard due to the robustness of our speech perception. Appreciating the importance of harmonicity for segregation, we empirically investigate the factors responsible for these models to produce a harmonic analysis of the mixture and subsequently track these harmonic patterns to eventually perform segregation.

The goal of our paper is to analyze the shortcomings of

10.21437/Interspeech.2022-11222

these E2E models to better understand the principles responsible for their success. We believe that this would help us - 1) understand the current limitations of these networks, 2) interrelate traditional Computational Auditory Scene Analysis (CASA) algorithms which rely on engineering [23,24], biological [25,26], or pitch tracking [27–29] based solutions with E2E networks, and 3) improve these networks by training them to be invariant to these shortcomings. We observe that-

- Speech Segregation is more reliant on cues from the lower harmonics of speech than the higher harmonics.
- E2E source segregation models are very sensitive to a discontinuous harmonic space. These networks fail when any intermediate harmonic is absent.
- Short discontinuities of 20ms in the audio cause a Causal ConvTasnet to perform incorrect assignment near the point of discontinuity
- Spectrogram-based models trained to generate masks directly on the mixture spectrogram are more robust to harmonic deformations and temporal discontinuities.

2. Experiments

We perform our analysis on ConvTasnet and DPT-Net trained to segregate 2 speakers. First, we investigate whether the models attend to the higher harmonics which are more prone to overlapping or to the lower harmonics which generally contain more energy. We then analyze the importance of continuity in the harmonic structure for grouping the channels of each estimated speaker. Next, we investigate the importance of temporal continuity to analyze how the models approach the assignment problem of deciding which estimated speaker should be assigned to which channel. Lastly, we contrast these results with speech-segregation networks trained directly on spectrograms [30].

2.1. Datasets and Evaluation Metrics

We perform our analysis on ConvTasnet and DPT-Net trained on the WSJ-2-Mix data [4]. We report experiments on both, mixtures of speech from the WSJ-2-Mix test-set and mixtures of non-overlapping alternating tones which are more controllable.

We demonstrate the network's behaviour visually by generating mixtures of two alternating, non-overlapping harmonic tones, where each tone has N harmonics and is given by $x_i(t) = \sum_{k=1}^{N} a_k \sin(2\pi k f_0^{(i)} t), i \in \{1, 2\}$. A mixture of these tones containing all the N harmonics can be successfully segregated by the networks. It has been shown in previous work that ConvTasnet relies very strongly on the harmonic structure [1]. Input stimuli of such nature allow us to focus on this harmonic structure and investigate how the network analyses it, by removing certain harmonics such as in Sec. 2.2.1 and Sec 2.2.2 and precisely inserting discontinuities in Sec 2.2.3. To perform our analysis on speech, we filter the validation set of the WSJ-2-Mix dataset using low-pass, high-pass or band-stop-filters to analyse the network's performance when missing the lowest, highest or intermediate harmonics. We evaluate the ConvTasnet model using the SDR.

2.2. Results

2.2.1. Model Performance on Low-Passed and High-Passed Filtered Speech

We investigate whether a network trained on natural speech attends more to the lower or higher harmonics for segregation. We low-pass (LP) filter the WSJ-2-mix dataset with cut-off frequencies at 300Hz, 700Hz, and 1200Hz and evaluate a model

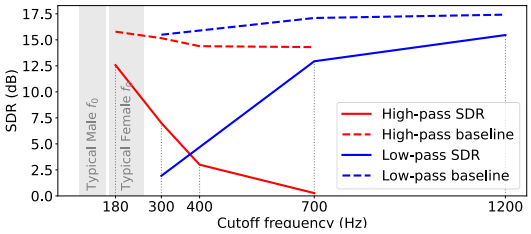

Figure 3: *Performance of ConvTasnet on either LP (blue) or HP (red) filtered speech for different cutoff frequencies, compared to the performance of networks trained on filtered data.*

with this filtered data to analyze if the network can detect the harmonic structure necessary for segregation [1] in the absence of the higher harmonics. Similarly, we also high-pass (HP) filter the dataset, at cut-off frequencies 180Hz, 300Hz, 400Hz and 700Hz and evaluate the model on this filtered data. We baseline these results against networks trained and evaluated on the corresponding filtered dataset. An alternate baseline is to segregate natural speech using the model trained on natural speech, and filtering the estimated sources. In both cases, we observe performances of around 15db and thus only report the former.

These results are illustrated in Fig 3, where it is evident that the network is very sensitive to speech missing its lowest harmonics (e.g. telephonic speech). The network is more robust to the LP filtered data, indicating that the model relies on cues in the lower harmonics more than in the higher harmonics to characterize the incoming mixture.

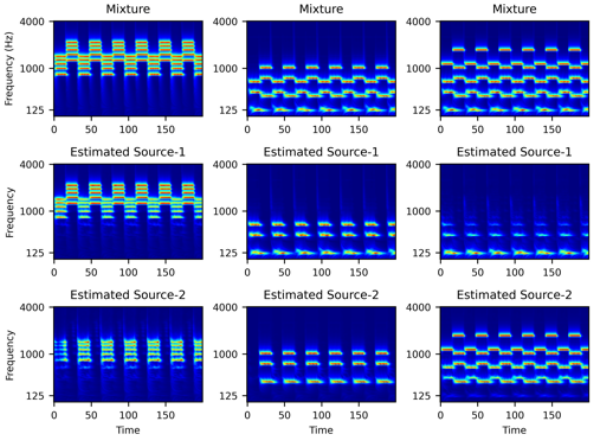

Figure 4: *Conv-Tanet fails to segregate mixtures of higher harmonics (left columns) but successfully segregates mixtures of the first three harmonics (centre columns). It fails to segregate mixtures of the first three and fifth harmonics (right columns)*

These results are corroborated by our observations in Fig. 4 on mixtures of two alternating tones with F0 as 117Hz and 201Hz. The left column indicates the model's inability to segregate a mixture of tones with the harmonics: $4, \cdots, 8$ and the middle column indicates the model's ability to segregate a mixture with the lowest three harmonics.

2.2.2. Model Performance on Band-Stopped Filtered Speech

We investigate the network's sensitivity to discontinuities in the harmonic structure. We generate speech missing a few inter-

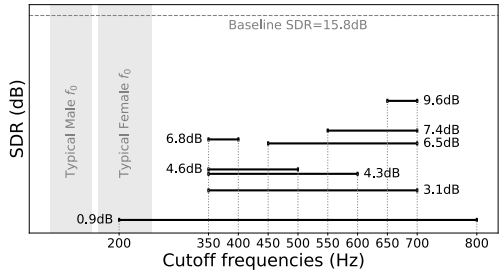

Figure 5: *Performance of ConvTasnet on band-stoped data. Each filter is a horizontal line representing the frequencies in the stop-band, and the corresponding SDR.*

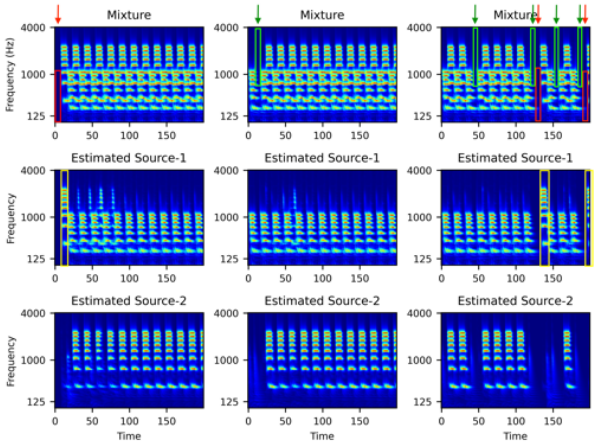

Figure 6: *Discontinuities are inserted into a mixture of alternating tones (upper rows). Errors in segregation can be observed when these discontinuities mute the lower frequency tone (yellow boxes in the middle rows).*

mediate harmonics by applying a band-stop filter to the mixture before doing the segregation. We perform this evaluation with a set of 8 band-stop filters, with stop-band frequencies ranging from 200Hz-800Hz, to 350Hz-400Hz. We report the results in Fig. 5 in which each horizontal line represents a band-stop filter by its low and high cutoff frequencies, and reports the corresponding SDR. As in 2.2.1, multiple baselines can be considered which all resulted in performances above 15dB. For clarity, we only report the ConvTasnet native performance of 15.8dB in Fig. 5. We repeat this study on the mixture of tones in Fig 4. While the network can segregate a mixture of only the first 3 harmonics (2^{nd} column), it fails when we introduce the 5^{th} harmonic (3^{rd} column). While doing so brings extra information, it is perceived as a missing 4^{th} harmonic which challenges the network's expectations. Unsure of the harmonic structure's position, the network fails and segregates both tones into the same source, indicating its dependence to a continuous harmonic structure.

2.2.3. Performance on Mixtures with Temporal Discontinuities

We study the temporal consistency of the network's output by introducing short silences in the data. To visualize the quality of segregation at the point of discontinuity, we use a causal ConvTasnet for these experiments. We create mixtures of two alternating and insert silence ranging from 15ms to 100ms since

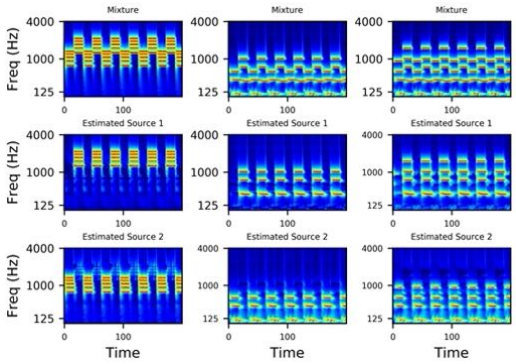

Figure 7: *A spectrogram-based network can successfully segregate mixtures of the higher (left column), first three (centre column) and first three and fifth (right column) harmonics*

such silence is usually present in speech, on which the network is trained. Fig 6 shows our results when tone period is 62ms (i.e. 31ms of energy) and the inserted silence (discontinuities) is 31ms in duration. This silence is inserted to the mixture at the points of the arrows in the upper row. The red (respectively green) arrows indicate that the silence has muted the low (respectively high) frequency tone. The corresponding red and green arrows in the mixture spectrograms (upper-rows) indicate the missing tone. We observe that the inserted discontinuity leads to an error in segregation, denoted by the yellow boxes in the estimated spectrograms. These errors are localized to the neighborhood of discontinuity, after which the network begins to segregate the sources correctly. More interestingly, these errors only arise when the low frequency tone is muted (red boxes in the mixture spectrogram). Discontinuities that arise from muting the higher frequency tones do not seem to cause an error in segregation. This observation is consistent across tones of periods ranging from 30ms to 125ms and the duration of the discontinuities as transient as 10ms, repeated using 10 random seeds. This behavior is discussed in Sec 2.3.

2.2.4. Comparison of E2E Models with Spectrogram Based Models

We analyze the behavior of spectrogram-based speech segregation models on tones with the same harmonic and temporal deformations. We train the separation sub-network of ConvTasnet in Fig 1 to directly generate masks on the spectrogram. This network is trained to optimize the L2 norm between the estimated and true spectrograms using PIT and is also trained on the WSJ-2-mix dataset. It performs equivalently to a state-of-art spectrogram-based model [8] on natural speech. As illustrated in Fig 7, when presented with the same stimuli as that in Fig 4, we observe that the spectrogram based model is not sensitive to missing harmonics tests. It can successfully segregate mixtures of just the higher harmonics (left column), lower harmonics (centre column), and mixtures with missing harmonics (right column). We also analyze the sensitivity of spectrogram-based models to temporal discontinuities and present the model with the same stimuli discussed in Sec 2.2.3. Fig 8 illustrates that the spectrogram based-model is robust to these deformations and can successfully segregate mixtures with periods of silence. Our results indicate that spectrogram-based models are significantly more robust to deformations in the evaluation data, in-spite of being trained on the same dataset.

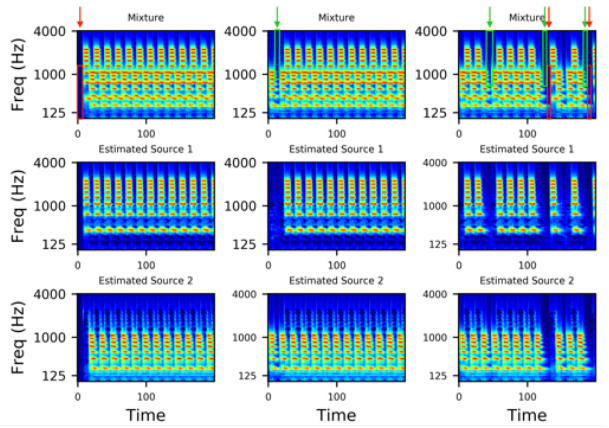

Figure 8: *A spectrogram-based model can successfully segregate mixtures with inserted silences which are denoted by red and green arrows in the upper rows.*

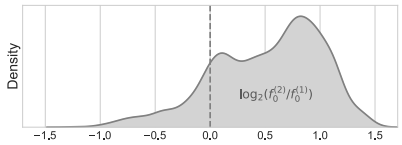

Figure 9: *Density over evaluation mixtures of the ratio* $\log_2(f_0^{(2)}/f_0^{(1)})$ *where* $f_0^{(1)}$ *and* $f_0^{(2)}$ *are the average F0s of ConvTasnet's output channels 1 and 2 respectively.*

2.3. Discussion

We have demonstrated that ConvTasnet is very sensitive to transformations on the harmonic structure. Removal of the lower harmonics (HP filtering), higher (LP filtering) or intermediate harmonics (band-stop filtering) often occur in the *wild*. These transformations result in an imperceptible deformation to humans but have a catastrophic impact on the network. We repeat our experiments on DPT-Net and observe similar results.

2.3.1. Sensitivity of Networks to Deformations in Harmonics

Section 2.2.1 indicates that E2E segregation networks cue on to the lower harmonics to characterize the harmonic structure. A possible explanation for this is that the lower harmonics have more energy and are often well separated, even in a mixture. Computing the F0 from the higher harmonics is less trivial. For example, energy at 250Hz could either be the F0 of a female speaker, or the second harmonic of a male speaker whose F0 is 125Hz. However, energy at 900Hz could be the 4^{th} or 5^{th} harmonic of a female speaking at respectively 225 or 180Hz, or the 6^{th}, 7^{th}, ..., or 10^{th} harmonic of a male speaking at respectively 150Hz, 128Hz, ..., or 90Hz. However, the collection of higher harmonics are perceptually important since humans can segregate speech composed of just the higher harmonics and compute the F0 [31]. Our experiments in Sec 2.2.2 show that even removing a small frequency band from $350-400$Hz or $650-700$Hz considerably degrades the network's performance, from above 15dB to 6.8dB and 9.6dB respectively. This effect is stronger when removing lower frequencies than higher frequencies, and when increasing the range of the stop-band.

2.3.2. Sensitivity of Networks to Temporal Discontinuities

Our experiments in Sec 2.2.3 demonstrate that E2E models quickly lose track of the speaker identities after short silences. This may be due to the difficulty of the assignment problem: even after performing speaker grouping, the model needs to decide at every time frame which speaker it should assign to a given output channel. This has been addressed in the past as the label permutation problem - networks are now trained using utterance-level permutation invariant training (U-PIT) [8, 32], which makes the loss invariant to the order of speakers in the output. However, although the chosen order does not affect the loss function, the network still needs to assign each speaker to a channel in a way that is stable over time. In feed-forward networks, this is decided without any information about past assignments. An option is to compute a statistic that can reliably order the speakers in a temporally consistent way. We show in Fig. 9 the distribution of $\log_2(f_0^{(2)}/f_0^{(1)})$ where $f_0^{(i)}$ is the average F0 estimated from the output channel i. This distribution is computed using the WSJ-2-mix test-set. It is strongly biased towards positive values, with one peak just above 0 and one near 1, corresponding to mixtures of speakers with the same or different genders respectively. We observe that for 86% of the mixtures, the lower frequency speaker is assigned to channel 1. The network seems to solve the assignment issue by ordering the speakers' fundamental frequencies. Interestingly, this learned trait is in alignment to pre-defined conventions used in early segregation models [33], without it being explicitly constrained during training. When unsure, the networks seem to assign almost all the energy to the channel 1, explaining the behavior at points of temporal discontinuity. Thus, they rely on the self-consistency of the speakers to perform segregation.

3. Conclusion

Our work investigates the performance of E2E speech segregation networks to data with harmonic deformations and temporal discontinuities. Given their poor performance for speech missing the lower and intermediate harmonics, we believe that these models strongly rely on cues from the lower harmonics to detect the harmonic structure.

The network's sensitivity to silences in the mixtures indicates that it enforces consistency of its output over a time window much shorter than its perceptive field. The network assigns speakers to different channels by comparing short-term features (*e.g.* pitch) and is only consistent over time because of the data's consistency.

We also demonstrate that spectrogram based models are more robust to the above deformations, indicating that ConvTasnet's vulnerabilities are not a result of a dataset that prevents *out-of-distribution* generalization but rather a consequence of the learned time-frequency representation. Given that the performance of E2E models have surpassed human performance in ideal conditions, their next challenge is to generalize to speech with such deformations. We believe that using transfer learning [34] or expert-knowledge in designing representations that are more invariant to common deformations, and data-augmentation during training may help bridge this gap in robustness.

4. Acknowledgements

This work was supported by NSF grant #1764010 and an AFOSR grant. The authors declare no conflict of interests.

5. References

[1] R. Parikh, I. Kavalerov, C. Espy-Wilson, and S. Shamma, "Harmonicity plays a critical role in dnn based versus in biologically-inspired monaural speech segregation systems," *arXiv preprint arXiv:2203.04420*, 2022.

[2] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[3] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," *arXiv preprint arXiv:2007.13975*, 2020.

[4] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 31–35.

[5] X.-L. Zhang and D. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 5, pp. 967–977, 2016.

[6] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[7] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 46–50.

[8] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[9] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[10] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," *arXiv preprint arXiv:1607.02173*, 2016.

[11] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 246–250.

[12] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 708–712.

[13] Z. Shi, H. Lin, L. Liu, R. Liu, J. Han, and A. Shi, "Deep attention gated dilated temporal convolutional networks with intra-parallel convolutional modules for end-to-end monaural speech separation." in *Interspeech*, 2019, pp. 3183–3187.

[14] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7092–7096.

[15] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. Le Roux, and J. R. Hershey, "Universal sound separation," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 175–179.

[16] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," *arXiv preprint arXiv:1911.13254*, 2019.

[17] S. Wisdom, H. Erdogan, D. P. W. Ellis, R. Serizel, N. Turpault, E. Fonseca, J. Salamon, P. Seetharaman, and J. R. Hershey, "What's all the fuss about free universal sound separation data?" in *in preparation*, 2020.

[18] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, New York City, United States, October 2019. [Online]. Available: https://hal.inria.fr/hal-02160855

[19] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[20] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[21] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[22] R. Parikh, "Demystifying monaural speech segregation models," Ph.D. dissertation, 2022.

[23] R. J. Weiss and D. P. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Computer Speech & Language*, vol. 24, no. 1, pp. 16–29, 2010.

[24] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Computer Speech & Language*, vol. 24, no. 1, pp. 1–15, 2010.

[25] M. Elhilali and S. A. Shamma, "A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation," *The Journal of the Acoustical Society of America*, vol. 124, no. 6, pp. 3751–3771, 2008.

[26] L. Krishnan, M. Elhilali, and S. Shamma, "Segregating complex sound sources through temporal coherence," *PLoS computational biology*, vol. 10, no. 12, p. e1003985, 2014.

[27] S. Vishnubhotla and C. Y. Espy-Wilson, "An algorithm for speech segregation of co-channel speech," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 109–112.

[28] M. Stark, M. Wohlmayr, and F. Pernkopf, "Source–filter-based single-channel speech separation using pitch information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 242–255, 2010.

[29] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE transactions on neural networks*, vol. 10, no. 3, pp. 684–697, 1999.

[30] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[31] S. Shamma and D. Klein, "The case of the missing pitch templates: how harmonic templates emerge in the early auditory system," *The Journal of the Acoustical Society of America*, vol. 107, no. 5, pp. 2631–2644, 2000.

[32] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[33] C. Weng, D. Yu, M. L. Seltzer, and J. Droppo, "Deep neural networks for single-channel multi-talker speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1670–1679, 2015.

[34] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.